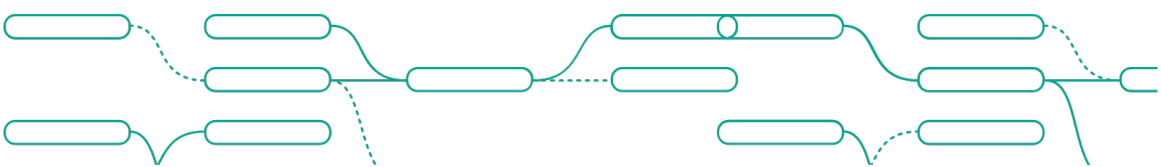
Arenadata Catalog

Высокоуровневая архитектура развертывания Arenadata Catalog AI Стюард данных 1.2

Москва 2026



Arenadata Catalog

Оглавление

Архитектура модуля AI Стюард данных **4**

Требования к вычислительной среде **5**

PROD-среда 5

TEST-среда 5

DEV-среда 5

Развертывание ADC AI Стюард данных для среды демонстрации, апробации 6

Установка компонентов на несколько серверов 6

Требования к программному обеспечению пользователя **7**

Требования к среде резервного копирования **7**

Рекомендации по моделям для AI Стюард данных **8**

Языковая модель 8

Модуль Arenadata Catalog AI Стюард данных поддерживает контейнерное развёртывание с использованием Docker Compose и установку компонентов непосредственно на физические или виртуальные серверы. Является дополнительным модулем системы и лицензируется отдельно. Поставляется отдельным пакетом и не входит в базовый состав дистрибутива Arenadata Catalog, предоставляемый заказчику.

Состав пакета дистрибутива ADC AI Стюард данных приведён в таблице 1.

Таблица 1. Компоненты продукта ADC AI Стюард данных

Функциональный блок	Компоненты продукта
ИИ-агенты, планировщик задач и REST API	AI Data Steward Service
Доступ сторонних ИИ-агентов к Arenadata Catalog по протоколу MCP	MCP Server

Компоненты Arenadata Catalog, необходимые для работы модуля, приведены в таблице 2. Модуль не запускает собственных экземпляров этих компонентов и использует те, что развёрнуты в составе базовой системы.

Таблица 2. Компоненты Arenadata Catalog, необходимые ADC AI Стюард данных

Функциональный блок	Компонент среды функционирования
Сервер приложения Http/API	Arenadata DC Platform
Служебная база данных	PostgreSQL 14*
Служебный брокер сообщений	Apache Kafka и Apache ZooKeeper в составе Confluent Platform 7.7.0
Сервис захвата метаданных	Arenadata DC Ingestion Framework**
Веб-интерфейс	Arenadata Catalog UI***

* Модуль не разворачивает отдельный сервер базы данных. При установке в служебной СУБД Arenadata Catalog создаётся отдельная база данных модуля.
** Требуется для запуска агентов по расписанию. При запуске агентов вручную или из внешних систем не обязателен.
*** Требуется для управления запусками агентов и просмотра результатов из интерфейса каталога.

Внешние компоненты, приобретаемые и разворачиваемые заказчиком отдельно, приведены в таблице 3.

Таблица 3. Внешние компоненты для среды функционирования ADC AI Стюард данных

Компонент среды функционирования	Примечание
Большая языковая модель (LLM) с API, совместимым с OpenAI	Поставляется заказчиком. Рекомендации по модели и оборудованию приведены в разделе «Рекомендации по моделям для AI Стюард данных»

Архитектура модуля AI Стюард данных

Возможна установка с использованием Docker Compose для контейнерной архитектуры либо установка каждого компонента по отдельности.
Модуль встроен в Arenadata Catalog и не является самостоятельным приложением с собственным интерфейсом. Пользователь работает с агентами из интерфейса каталога, а все обращения к модулю проходят через сервер приложения Arenadata Catalog. Прямых обращений пользователей к модулю в обход сервера приложения не предусмотрено.
При контейнерном развёртывании AI Data Steward Service работает в одном контейнере и совмещает в себе REST API для управления запусками агентов, планировщик, распределяющий задачи между рабочими потоками и ИИ-агенты.

Сценарии запуска агентов:

- **Из интерфейса каталога.** Пользователь запускает агента на выбранных объектах, сервер приложения передаёт запрос в REST API модуля и отображает ход выполнения и результаты.
- **По расписанию.** В Arenadata DC Ingestion Framework настраивается конвейер ADC AI Стюард данных. Он конфигурируется на отдельной вкладке в интерфейсе каталога, по образцу конвейера переиндексации, и по расписанию создаёт запуск агента с отобранными объектами.

Для обработки SQL-кода из Git и S3 сервису необходим доступ к выбранному репозиторию или бакету с правами чтения.

Дополнительно в составе модуля поставляется MCP Server, открывающий сторонним ИИ-агентам доступ к объектам Arenadata Catalog по протоколу MCP. Сторонний агент передаёт на MCP Server JWT-токен пользователя, от имени которого он действует, и все обращения к каталогу выполняются с правами этого пользователя.

Модуль обращается к серверу приложения Arenadata Catalog под учётной записью служебного пользователя-бота, создаваемой при установке.

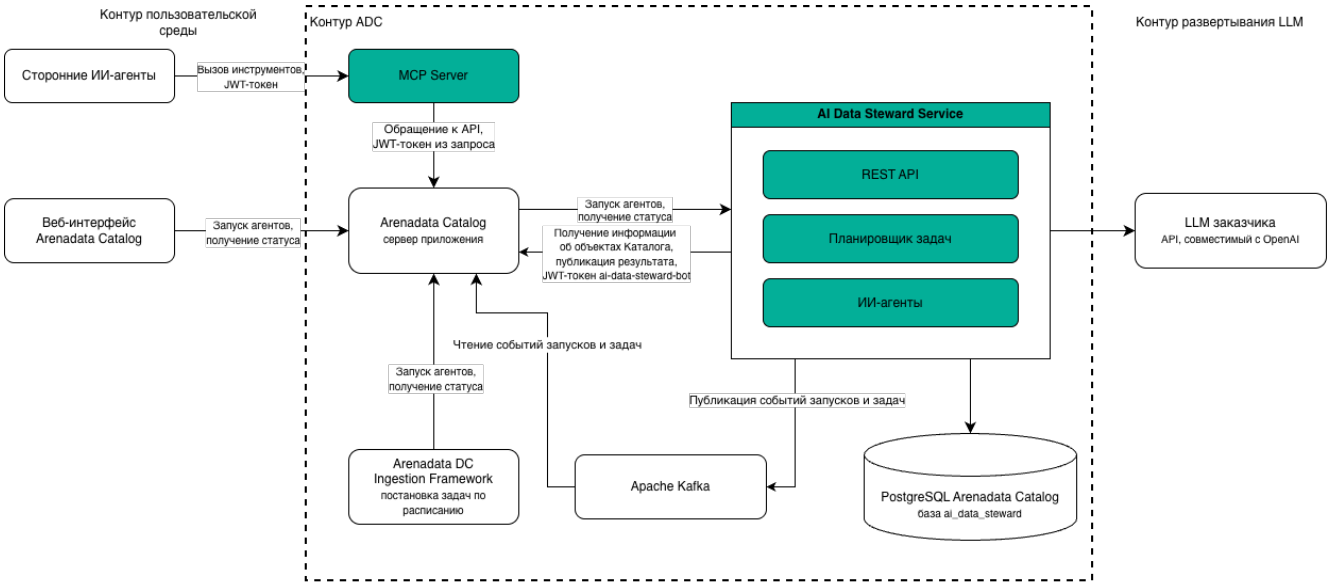


Рисунок 1. Высокоуровневая архитектура продукта Arenadata Catalog AI Стюард данных.

Arenadata Catalog

Требования к вычислительной среде

Модуль разворачивается в дополнение к Arenadata Catalog, поэтому приведённые ниже требования являются дополнительными к требованиям базовой системы, изложенным в документе «Высокоуровневая архитектура развертывания Arenadata Catalog».

PROD-среда

Требования к вычислительной среде для PROD-среды рассчитываются индивидуально и зависят от объёма каталога, числа обрабатываемых объектов и числа одновременно выполняемых агентов.

Минимальные требования к серверу модуля:
- 4 ядра CPU x64, не менее 2 ГГц;
- 8 ГБ оперативной памяти при значении параметра `RUNNER_MAX_CONCURRENT_TASKS` по умолчанию (16 одновременно выполняемых задач);
- При увеличении параллелизма объём памяти масштабируется пропорционально;
- 100 ГБ полезного дискового пространства HDD/SSD (с минимальной скоростью чтения 150 МБ/с).

Дополнительные требования к серверу базы данных Arenadata Catalog:
- не менее 20 ГБ полезного дискового пространства SSD дополнительно к требованиям базовой системы.

Объём служебной базы данных растёт с историей запусков. Автоматической очистки не предусмотрено: устаревшие запуски удаляются администратором через API модуля.

TEST-среда

Допускается развёртывание модуля на сервере приложения Arenadata Catalog.
Требования к серверу модуля:
- 2 ядра CPU x64, не менее 2 ГГц;
- 8 ГБ оперативной памяти;
- 50 ГБ полезного дискового пространства HDD/SSD (с минимальной скоростью чтения 150 МБ/с).

Дополнительные требования к серверу базы данных:
- Не менее 10 ГБ полезного дискового пространства SSD дополнительно к требованиям базовой системы.

При необходимости проведения нагрузочного тестирования рекомендуется использовать вычислительную среду, идентичную по ресурсам PROD-среде.

DEV-среда

Минимальные технические требования к вычислительной инфраструктуре для контура разработки.
Требования к серверу модуля:
- 2 ядра CPU x64, не менее 2 ГГц;
- 4 ГБ оперативной памяти;
- 50 ГБ полезного дискового пространства HDD/SSD (с минимальной

- скоростью чтения 150 МБ/с).

Дополнительные требования к серверу базы данных:

- Не менее 10 ГБ полезного дискового пространства SSD дополнительно к требованиям базовой системы.

Развертывание ADC AI Стюард данных для среды демонстрации, апробации

Возможна установка модуля на тот же сервер, где развёрнут Arenadata Catalog, при следующих минимальных требованиях:

- 2 ядра CPU x64, не менее 2 ГГц;
- 4 ГБ оперативной памяти;
- 50 ГБ полезного дискового пространства HDD/SSD (с минимальной скоростью чтения 150 МБ/с).

Требования к серверу инференса языковой модели приведены в разделе «Рекомендации по моделям для AI Стюард данных» и в среде демонстрации не снижаются.

Установка компонентов на несколько серверов

Раздел описывает размещение компонентов на отдельных серверах, а не горизонтальное масштабирование: AI Data Steward Service работает в одном экземпляре, запуск нескольких экземпляров одного сервиса не поддерживается.
Требования к серверам в случае распределённой установки зависят от многих факторов и являются предметом дополнительного обсуждения. В качестве рекомендаций используйте приведённые ниже характеристики для промышленной PROD-среды.

AI Data Steward Service

- 4 ядра, не менее 2 ГГц;
- 8 ГБ оперативной памяти;
- 100 ГБ SSD.

MCP Server

- 2 ядра, не менее 2 ГГц;
- 2 ГБ оперативной памяти;
- 10 ГБ SSD.

Сервер инференса языковой модели

- 16-32 ядра, не менее 2 ГГц;
- 128-256 ГБ оперативной памяти;
- NVMe SSD от 240 ГБ;
- Графический ускоритель согласно разделу «Рекомендации по моделям для AI Стюард данных».

Требования к серверам компонентов базовой системы Arenadata Catalog приведены в документе «Высокоуровневая архитектура развертывания Arenadata Catalog».

Требования к программному обеспечению пользователя

Вся функциональность модуля доступна пользователю через веб-браузер в интерфейсе Arenadata Catalog. Собственных требований к браузеру модуль не предъявляет. Поддерживаемые дистрибутивы и версии браузеров приведены в документе «Высокоуровневая архитектура развертывания Arenadata Catalog».

Требования к среде резервного копирования

Требования к вычислительным ресурсам для резервного копирования и восстановления варьируются в зависимости от различных факторов: политик выполнения резервного копирования, наличия инструментов для выполнения операций, требований к восстановлению данных (RPO, RTO) и т.д. Требования к среде резервного копирования модуля являются предметом дополнительного обсуждения.
Служебная БД модуля ai_data_steward размещается в общей СУБД Arenadata Catalog и резервируется отдельным дампом PostgreSQL в рамках общего регламента резервного копирования ADC. Состояние запусков и задач модуля восстанавливается из резервной копии его служебной БД, а результаты работы агентов, сохранённые в каталоге, — из резервной копии БД каталога. Порядок копирования и восстановления описан в инструкции по установке и эксплуатации ADC.

Рекомендации по моделям для AI Стюард данных

Языковая модель

Приведённые ниже требования к оборудованию относятся к случаю, когда заказчик разворачивает модель самостоятельно. При использовании внешнего сервиса с API, совместимым с OpenAI, они не применяются.
Рекомендуется Qwen3.5-35B-A3B, с весами в формате FP8, сервер инференса vLLM 0.17 или новее.
Рекомендуемое оборудование и ожидаемая скорость инференса приведены в таблице 4.

Таблица 4. Рекомендуемое оборудование для языковой модели.

Конфигурация	Контекст запроса	Одновременных запросов	Скорость генерации
1 × H100 80 ГБ	32K	около 90	около 51 токен/с на запрос при 10 одновременных запросах; одиночный запрос – до 210 токенов/с
1 × H100 80 ГБ	262K	около 11	при полной загрузке оценочно 8-15 токенов/с на запрос; одиночный запрос – 143 токена/с

Требования к остальному оборудованию сервера инференса, помимо графического ускорителя, приведены в таблице 5.

Таблица 5. Требования к серверу инференса помимо графического ускорителя.

Компонент	Рекомендация
CPU	16-32 ядер: токенизация запросов и обслуживание API выполняются на процессоре
Оперативная память	128-256 ГБ
Диск	NVMe SSD от 240 ГБ: веса модели занимают около 40 ГБ, дополнительно требуется место под кэш и журналы. Быстрый диск сокращает время загрузки модели при старте

Примечания к таблицам 4 и 5:

- Приведённые значения являются приблизительной оценкой по данным публичных замеров, а не результатом нагрузочного тестирования в составе Arenadata Catalog.
- Число одновременных запросов указано для случая, когда все запросы одновременно удерживают полный контекст. Память расходуется по фактическому размеру контекста каждого запроса, поэтому при реальной нагрузке одновременно обслуживается больше запросов.
- Значения CPU, оперативной памяти и диска соответствуют типовой комплектации серверов с ускорителем H100 в облачных предложениях.
- Для работы модуля необходимы OpenAI-совместимый Chat Completions API, контекстное окно от 32 тыс. токенов, поддержка вызова инструментов (tool calling) в автоматическом режиме (tool_choice=auto) и авторизация по Bearer-токену. Модуль поддерживает несколько способов получения структурированного ответа от LLM. Режим выбирается в соответствии с возможностями модели и API провайдера. Выбор модели заказчиком не ограничивается; приведённая рекомендация обеспечивает заявленное качество работы агентов.

Документ предоставлен исключительно в информационных целях, и его содержание может быть изменено без предварительного уведомления.

Все торговые марки, знаки обслуживания, зарегистрированные торговые знаки или зарегистрированные знаки обслуживания являются собственностью их законных владельцев.